2Я

УДК 519.688 + 336.144

А.И. Якупов
Ижевский государственный технический университет, г. Ижевск, Российская Федерация
yakupov_aidar@mail.ru

Применение деревьев решений для моделирования кредитоспособности клиентов коммерческого банка

В настоящее время у российских банков остро стоит задача в оптимизации процесса выдачи кредитов как физическим, так и юридическим лицам. Это позволяет сделать скоринг. В данной статье описывается применение деревьев решений для построения скоринговой модели.

Основу финансовой системы современной России составляют банки. Поэтому от успешного развития банковской деятельности зависит устойчивость как финансовой системы, так и всей экономики. Российская банковская система находится на этапе интенсивных рыночных преобразований. В условиях острой конкуренции и концентрации банковского капитала важнейшей задачей является оптимизация кредитования как физических, так и юридических лиц. На сегодняшний день российские банки предоставляют огромное количество «кредитных продуктов» на различные цели. Кредитование является основной важнейшей деятельностью любого банка. Но, к сожалению, банки не используют весь свой имеющийся потенциал для кредитования. Это связано с тем, что кредитование связано с большим количеством рисков, в первую очередь с риском невозврата выданных средств. Поэтому при принятии решения о выдаче кредита банк всячески перестраховывается, а это непосредственно влияет на то, что не все потенциальные «честные» заемщики могут соответствовать требованиям, предъявляемым банком к заемщику. Также даже если заемщик и соответствует требованиям банка, он может просто отказаться кредитоваться из-за того, что процедура кредитования достаточно сложная и занимает много времени. Соответственно количество заемщиков у банка уменьшается, уменьшается количество выданных кредитов, поэтому происходит в некотором роде застой денежных средств и, соответственно, банк использует свой капитал не оптимально, вследствие чего теряет прибыль.

Во многих развитых странах мира эта проблема решена с помощью так называемых скоринговых систем, которые позволяют оптимизировать процесс кредитования. Буквальный перевод «scoring» – «выигрыш». Скоринг по существу является методом классификации всей интересующей нас популяции на различные группы, когда нам неизвестна характеристика, которая разделяет эти группы (вернет клиент кредит или нет), но зато известны другие характеристики, связанные с интересующей нас популяцией. В 1941 г. Дэвид Дюран впервые применил данную методику к классификации кредитов на «плохие» и «хорошие».

В настоящее время, в зависимости от задач анализа заемщика, кредитный скоринг включает application-скоринг – оценку кредитоспособности претендентов на получение кредита (скоринг по анкетным данным используется в первую очередь),

behavioral-скоринг – оценка вероятности возврата выданных кредитов (поведенческий анализ), а также collection-скоринг – оценка возможности полного либо частичного возврата кредита при нарушении сроков погашения задолженности (расчет рисков по портфелю) [1].

Скоринг представляет собой сложную математическую модель, которая позволяет классифицировать клиентов банка на различные группы, позволяя с определенной вероятностью отсеивать «плохих» клиентов. В целях построения модели сначала производится выборка клиентов кредитной организации, о которых уже известно, хорошими заемщиками они себя зарекомендовали или нет. Такая выборка называется «обучающей». Она может варьироваться от нескольких тысяч до сотни тысяч. Выборка подразделяется на две группы: «хорошие» и «плохие» риски. Это оправдано в том смысле, что банк при принятии решения о кредитовании на первом этапе выбирает из двух вариантов: давать кредит или не давать. Таким образом, скоринг представляет собой классификационную задачу, где, исходя из имеющейся информации, необходимо получить функцию, наиболее точно разделяющую выборку клиентов на «плохих» и «хороших».

Традиционными и наиболее распространенными являются линейные многофакторные регрессионные методы. Регрессионные методы плохо приспособлены для работы с переменными, выраженными в шкале наименований.

Логистическая регрессия позволяет преодолеть этот недостаток. В настоящее время логистическая регрессия широко применяется в скоринговых системах. Логистическая регрессия позволяет подразделять клиентов на несколько групп риска.

Все регрессионные методы чувствительны к корреляции между характеристиками, поэтому в модели не должно быть сильно коррелированных независимых переменных. Кроме того, регрессионные коэффициенты дают немного информации о механизме влияния рассматриваемых переменных на величину риска.

Дерево классификации представляют собой системы, которые разделяют клиентов на группы, внутри которых уровень риска одинаков и максимально отличается от уровня риска других групп.

Классификация выборки производится только на клиентах, которым дали кредит. Неизвестно, как бы повели себя клиенты, которым в кредите было отказано. Возможно, что какая-то часть оказалась бы приемлемыми заемщиками. Банкам следует фиксировать причины отказа и сохранять информацию. Это позволяет им восстанавливать первоначальную популяцию клиентов, обращавшихся за кредитом. С течением времени меняются и социально-экономические условия, влияющие на поведение людей. Поэтому скоринговые модели необходимо разрабатывать на выборке из последних клиентов, периодически проверять качество работы системы и, когда качество ухудшается, разрабатывать новую модель [2].

Важным направлением обработки данных являются рассуждения, основанные на предыдущем опыте [3]. Это методология, моделирующая нечеткий механизм размышлений, что сходно с процессом вывода заключений экспертами предметной области. Поля данных, используемых для объяснения и предсказания результата, становятся признаками ситуации. Число реальных событий должно быть достаточным для возможно более полного покрытия предметной области. Такие алгоритмы допускают представление информационных полей в цифровом виде, а также в виде лингвистических, булевых и дискретных переменных. В процессе поиска система использует либо некоторые из этих полей, либо все поля полностью, выполняя вычисления для объяснения или предсказания результата. Итоговое поле признаков

или любое другое поле, возникшее в результате моделирования взаимосвязей в полях исходных данных, может быть выражено в виде некоторого правила. Если данные неполные, алгоритм способен продолжить работу, извлекая наиболее подходящий результат. Подобные алгоритмы не предъявляют жестких требований к точности и полноте данных. Текстовые поля, содержащие качественную информацию, могут быть включены в процесс вывода обычным образом. Эти системы обеспечивают пользователя полезными инструкциями, содержащими оценку того, как исходные данные подводятся к итоговому решению. К алгоритмам анализа, основанным на правилах, следует отнести адаптивные системы нечеткого вывода и деревья решений. Метод деревьев решений отличается высокой скоростью обработки данных и обучения при сохранении свойств систем нечеткого логического вывода. Метод деревьев решений может применяться для целевой переменной, имеющей булев или категориальный тип. Такие переменные содержат значения, принадлежащие некоторому конечному множеству без определенного отношения порядка на нем.

Пусть целевая переменная соответствует некоторым классам, на которые разбито множество данных [3]. Требуется отыскать некоторое классифицирующее правило, позволяющее разбить множество данных на эти классы. В процессе поиска классифицирующего правила проводится перебор всех независимых переменных и отыскивается наиболее представительное правило на данном этапе. В обычных деревьях решений применяются предикаты вида $x \le w, x > w$. Данные разбиваются на две группы в соответствии со значением этого предиката. После этого процесс повторяется для каждой из этих групп до тех пор, пока получающиеся подгруппы содержат в себе представителей классов и включают в себя достаточно большое количество точек для того, чтобы статистически значимо быть разбитыми на меньшие подгруппы. В результате окончательное классифицирующее правило, построенное этим процессом, может быть представлено в виде бинарного дерева. Каждый узел этого дерева соответствует некоторому подмножеству данных и содержит найденное классифицирующее правило для этого подмножества.

Удобным для анализа свойством деревьев решений является представление данных в виде иерархической структуры. Компактное дерево проявляет картину влияния различных факторов, независимых переменных.

Метод классификации, основанный на деревьях решений, имеет в качестве преимуществ следующие свойства:
– быстрый процесс обучения;
– генерация правил в областях, где эксперту трудно формализовать свои знания;
– извлечение правил на естественном языке;
– интуитивно понятная классификационная модель;
– достаточно высокая точность прогноза, сопоставимая с другими методами;
– построение непараметрических моделей.

Эти положительные свойства приближают методологию деревьев решений к системам, основанным на нечеткой логике, выигрывая у них в быстроте процесса обучения.

Деревья решений – один из методов извлечения знаний из данных. Введем основные понятия из теории деревьев решений:
– объект – пример, шаблон, наблюдение, точка в пространстве атрибутов;
– атрибут – признак, независимая переменная, свойство;
– метка класса – зависимая переменная, целевая переменная, признак, определяющий класс объекта;

– узел – внутренний узел дерева, узел проверки;
– лист – конечный узел дерева, узел решения;
– проверка – условие в узле.

Деревья решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение. Под правилом понимается логическая конструкция, представленная в виде *if A then B* $(A \to B)$.

Пусть задано некоторое обучающее множество X, содержащее объекты, каждый из которых характеризуется m атрибутами и один из них указывает на принадлежность объекта к определенному классу. Это множество обозначим $X = \{\mathbf{x}^j, C_k^j\}, j = \overline{1,p}; k = \overline{1,K}$, где каждый элемент этого множества описывается атрибутами $\mathbf{x} = (x_i), i = \overline{1, m-1}$ и принадлежит одному из классов C_k. Количество примеров в множестве равно p является мощностью этого множества $|X|$. Через $\{C_k\}$ обозначается множество классов.

Каждое множество X на любом этапе разбиения характеризуется следующими показателями:

1) множество X содержит несколько объектов, относящихся к одному классу C_k. В этом случае множество X является листом, определяющим класс C_k;

2) множество X не содержит ни одного объекта (X =Ø). В данной ситуации необходимо возвратиться к предыдущему этапу разбиения;

3) множество X содержит объекты, относящиеся к разным классам. Такое множество является пригодным для разбиения на некоторые подмножества. Для этого выбирается одна из переменных и в соответствии с правилами $x \le w, x > w$ множество X разбивается на два подмножества. Этот процесс рекурсивно продолжается до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Данная процедура лежит в основе многих алгоритмов построения деревьев решений (метод разделения и захвата). Построение дерева решений происходит сверху вниз. Сначала создается корень дерева, затем потомки корня и т.д.

Поскольку все объекты были заранее отнесены к известным классам, такой процесс построения дерева решений называется обучением с учителем.

При построении деревьев решений необходимо решить следующие задачи:

а) выбор критерия атрибута, по которому пойдет разбиение;

б) остановка обучения;

в) отсечение ветвей.

Выбор критерия атрибута.

Для построения дерева на каждом внутреннем узле необходимо найти такое условие, которое бы разбивало множество, ассоциированное с этим узлом на подмножества. В качестве такой проверки должен быть выбран один из атрибутов. Выбранный атрибут должен разбить множество так, чтобы получаемые в итоге подмножества состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому, то есть количество объектов из других классов в каждом из этих множеств было как можно меньше.

Одним из способов выбора наиболее подходящего атрибута является применение теоретико-информационного критерия.

Задача заключается в построении иерархической классификационной модели в виде дерева из множества объектов $X = \{\mathbf{x}^j, C_k^j\}, j = \overline{1, p}; k = \overline{1, K}$. На первом шаге имеется только корень и исходное множество, ассоциированное с корнем.

Требуется разбить исходное множество на подмножества. Это можно сделать, выбрав один из атрибутов в качестве проверки. Тогда в результате разбиения получаются n (по числу значений атрибута) подмножеств и соответственно создаются n потомков корня, каждому из которых поставлено в соответствие свое подмножество, полученное при разбиении множества $X = \{\mathbf{x}^j, C_k^j\}, j = \overline{1, p}; k = \overline{1, K}$. Затем эта процедура рекурсивно применяется ко всем подмножествам (потомкам корня) и т.д. Любой из атрибутов можно использовать неограниченное количество раз при построении дерева.

Определим в качестве проверки t какой-либо атрибут, принимающий значения $x_i = (w_{ij}), i = \overline{1, m}; j = \overline{1, n}.$ Тогда разбиение X по проверке t дает соответствующие подмножества $X_j, j = \overline{1, n}$. Критерий выбора определяется информацией о том, каким образом классы распределены в множестве X и его подмножествах, получаемых при разбиении по t.

Обозначим $P_{iq}^k, k = \overline{1, K}; i = \overline{1, m}; q = \overline{1, n}$ вероятность принадлежности классу k по атрибуту i и q-му пороговому значению $x_i = (w_{ij}), i = \overline{1, m}; j = \overline{1, n}$, а P^k – вероятность попадания в класс k. В качестве меры среднего количества информации, необходимого для определения класса примера из множества X, берется энтропия Шеннона

$$H_X = -\sum_{k=1}^{K} P^k \log_2 P^k .$$

Разбиению множества X по проверке t соответствует выражение для энтропии

$$H_{iq} = -\sum_{k=1}^{K} P_{iq}^k \log_2 P_{iq}^k .$$

Критерием выбора является выражение

$$H_X - H_{iq} \to \max$$

или

$$H_{iq} \to \min .$$

Минимальное значение энтропии H_{iq} соответствует максимуму вероятности появления одного из классов. Выбранный атрибут и пороговое значение, минимизирующее H_{iq},

$$(i, q) = ArgMin\, H_{iq}$$

используются для проверки значения по этому атрибуту и дальнейшее движение по дереву производится в зависимости от полученного результата.

Данный алгоритм применяется к полученным подмножествам и позволяет продолжить рекурсивно процесс построения дерева до тех пор, пока в узле не окажутся примеры из одного класса. Если в процессе работы алгоритма получен узел, ассоциированный с пустым множеством (то есть ни один пример не попал в данный узел), то он помечается как лист, и в качестве решения листа выбирается наиболее часто встречающийся класс у непосредственного предка данного листа.

Для нахождения пороговых величин для числового атрибута значения $x_i^j, i=\overline{1,m}; j=\overline{1,p}$ сортируются по возрастанию и

$$w_{ij}=\frac{\left(x_i^j+x_i^{j+1}\right)}{2}, i=\overline{1,m}; j=\overline{1,p-1}$$

определяют порог, с которым должны сравниваться все значения атрибута. Следует отметить, что все числовые тесты являются бинарными, то есть делят узел дерева на две ветви.

Правила остановки разбиения узла.

1. Использование статистических методов для оценки целесообразности дальнейшего разбиения.

2. Ограничение глубины дерева. Остановить дальнейшее построение, если разбиение ведет к дереву с глубиной, превышающей заданное значение.

3. Разбиение должно быть нетривиальным, то есть получившиеся в результате узлы должны содержать не менее заданного количества примеров.

Правило отсечения ветвей дерева.

Предназначено для предотвращения сложных деревьев, трудных для понимания, которые имеют много узлов и ветвей.

Примем за точность распознавания дерева решений отношение правильно классифицированных объектов при обучении к общему количеству объектов из обучающего множества, а под ошибкой – количество неправильно классифицированных. Предположим, что известен способ оценки ошибки дерева, ветвей и листьев. Тогда возможно использовать следующее правило:

1 – построить дерево;

2 – отсечь или заменить поддеревом те ветви, которые не приведут к возрастанию ошибки.

Отсечение ветвей происходит снизу вверх, двигаясь с листьев дерева, отмечая узлы как листья, либо заменяя их поддеревом.

Вывод

В данной работе для построения модели взаимодействия с клиентами использован интеллектуальный метод обработки информации на основе деревьев решений.

Литература

1. Credit Scoring // Washington Post. – 2003. – December 11.
2. Воловник А.Д. Динамическое моделирование деятельности коммерческого банка. – Мурманск; Ижевск: Изд-во Кольского НЦ РАН, 2006.
3. Тененев В.А., Ворончак В.И. Решение задач классификации и аппроксимации с применением нечетких деревьев решений // Интеллектуальные системы в производстве. – 2005. – № 2.

А.І. Якупов

Застосування дерев рішень для моделювання кредитоспроможності клієнтів комерційного банку

У теперішній час перед російськими банками гостро постає задача оптимізації процесу видачі кредитів як фізичним, так і юридичним особам. Це дозволяє зробити скоринг. У даній статті описується застосування дерев рішень для побудування скорингової моделі.

A.I. Yakupov

The Decision Tree's Adaptation for Modeling of Commercial Bank Clients Solvency

Nowadays the main problem of the Russian banks which consist in crediting optimization both a natural person and a legal person. The “scoring” can help to decide this problem. The decision tree's adaptation for “scoring” model definition describe in this article.

Статья поступила в редакцию 02.07.2008.